

DS 4400

Machine Learning and Data Mining I  
Spring 2021

Alina Oprea  
Associate Professor  
Khoury College of Computer Science  
Northeastern University

January 26 2021

# Today's Outline

- Announcements
  - HW 1 is out, due on Friday, Feb 5
  - First numpy tutorial by Prabal M.
    - Thu, Jan 28, 5-6pm, Zoom link for office hours
- Probability review
  - Conditional probabilities
  - Bayes Theorem
- Linear algebra review
  - Matrix and vector operations
  - Transpose, inverse
  - Rank of a matrix

# Probability review

# Probability Resources

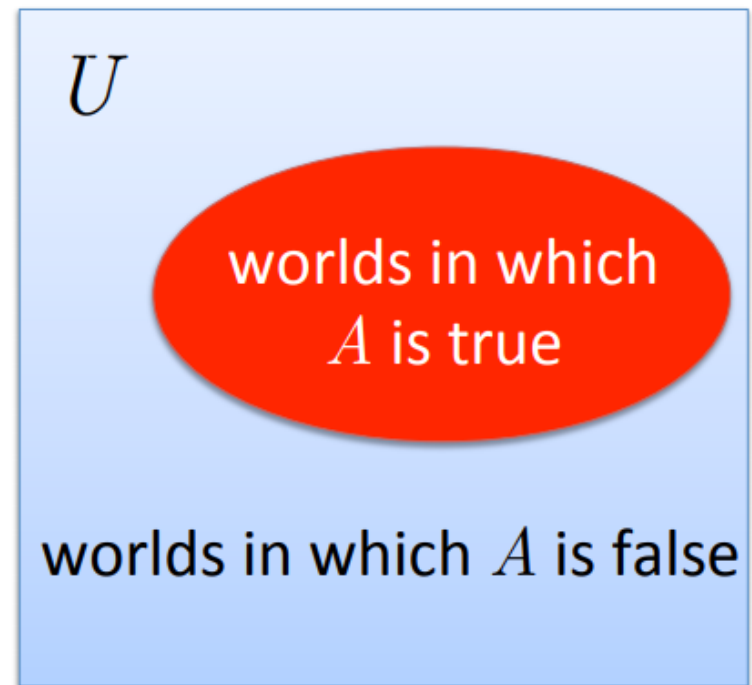
- [Review notes](#) from Stanford's machine learning class
- Sam Roweis's [probability review](#)
- David Blei's [probability review](#)
- Books:
  - Sheldon Ross, A First course in probability

# Discrete Random Variables

- Let  $A$  denote a random variable
  - $A$  represents an event that can take on certain values
  - Each value has an associated probability
- Examples of binary random variables:
  - $A$  = It will snow tomorrow
  - $B$  = The patient will recover
- $P(A)$  is “the fraction of possible worlds in which  $A$  is true”

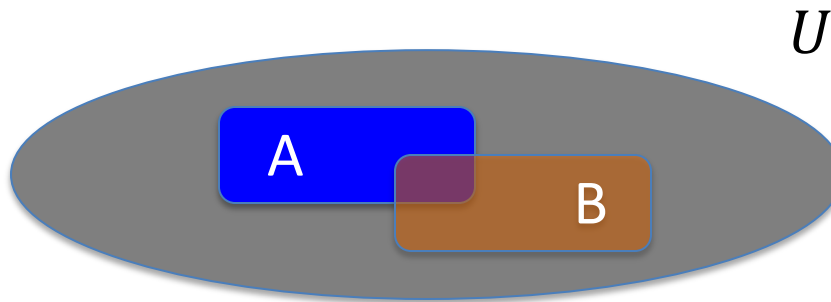
# Visualizing A

- Universe  $U$  is the event space of all possible worlds
  - Its area is 1
  - $P(U) = 1$
- $P(A) = \text{area of red oval}$
- Therefore:
$$P(A) + P(\neg A) = 1$$
$$P(\neg A) = 1 - P(A)$$



# Working with Probabilities

- $0 \leq P(A) \leq 1$
- $P(U) = 1; P(\Phi) = 0$
- $P(\neg A) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



**Union bound**

$$P(A \cup B) \leq P(A) + P(B)$$

# Examples discrete RV

- Bernoulli RV
  - $X$  is modelling a coin toss
  - Output: 1 (head) or 0 (tail)
  - $P[X=1] = p$ ;  $P[X=0] = 1-p$
- $Y$  is the number of points in a fair dice
  - $k \in \{1, \dots, 6\}$ ,  $P[Y = k] =$
  - $P[Y = \text{even}] =$



# Example discrete RV

- $Z$  is the sum of two fair dice
  - What is  $P[Z = k]$  for  $k \in \{2, \dots, 12\}$ ?
  - What is  $k$  for which this probability is maximum?

# Expectation

**Expectation** for discrete random variable  $X$

$$E[X] = \sum_v v \Pr[X = v]$$

Bernoulli:  $P[X=1] = p$ ;  $P[X=0] = 1-p$

# Expectation and variance

**Expectation** for discrete random variable  $X$

$$E[X] = \sum_v v \Pr[X = v]$$

## Properties

- $E[aX] = a E[X]$
- $E[X + Y] = E[X] + E[Y]$
- $E[f(X)] = \sum_v f(v) \Pr[X = v]$

## Variance

$$\text{Var}[X] \triangleq E[(X - E(X))^2]$$

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2, \end{aligned}$$

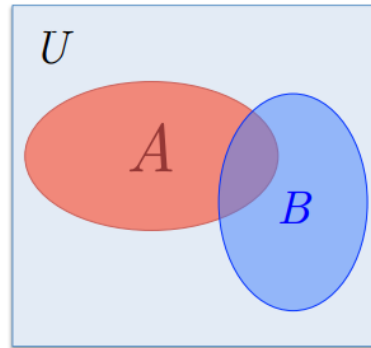
# Variance of Bernoulli

- **Variance:**  $\text{Var}[X] = E(X^2) - E^2(X)$

Bernoulli:  $P[X=1] = p$ ;  $P[X=0] = 1-p$

# Conditional Probability

- $P(A \mid B)$  = Fraction of worlds in which  $B$  is true that also have  $A$  true



What if we already know that  $B$  is true?

That knowledge changes the probability of  $A$

- Because we know we're in a world where  $B$  is true

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

Events  $A$  and  $B$  are **independent** if  $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$

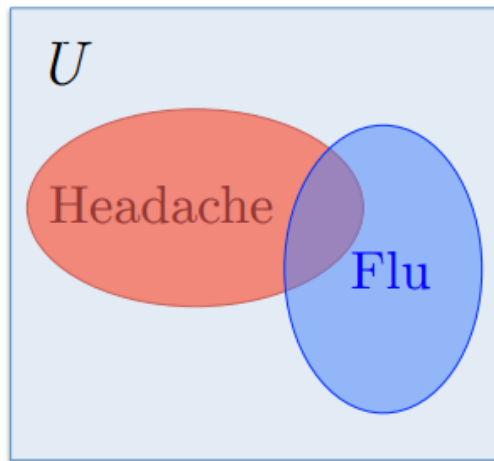
If  $A$  and  $B$  are independent

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[A]\Pr[B]}{\Pr[B]} = \Pr[A]$$

# Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$



$$P(\text{headache}) = 1/10$$

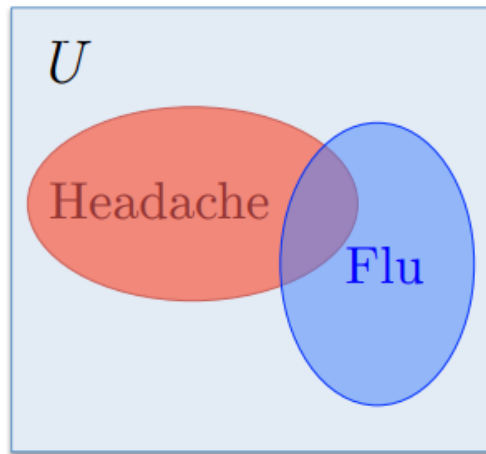
$$P(\text{flu}) = 1/40$$

$$P(\text{headache} \mid \text{flu}) = 1/2$$

“Headaches are rare and flu is rarer, but if you’re coming down with the flu there’s a 50-50 chance you’ll have a headache.”

# Inference from Conditional Probability

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$
$$P(A \wedge B) = P(A | B) \times P(B)$$



$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} | \text{flu}) = 1/2$$

One day you wake up with a headache.  
You think: “Drat! 50% of flus are  
associated with headaches so I must have  
a 50-50 chance of coming down with flu.”

Is this reasoning good?

# Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} \mid \text{flu}) = 1/2$$

Want to solve for:

$$P(\text{headache} \wedge \text{flu}) = ?$$

$$P(\text{flu} \mid \text{headache}) = ?$$

⋮



# Exercises

- Compute Expectation and Variance for dice rolling random variable  $X$ 
  - $P[X=k] = 1/6$  for  $k \in \{1, \dots, 6\}$
- Conditional probabilities

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} \mid \text{flu}) = 1/2$$

Want to solve for:

$$P(\text{headache} \wedge \text{flu}) = ?$$

$$P(\text{flu} \mid \text{headache}) = ?$$



# Inference from Conditional Probability

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}$$
$$P(A \wedge B) = P(A | B) \times P(B)$$

$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} | \text{flu}) = 1/2$$

Want to solve for:

$$P(\text{headache} \wedge \text{flu}) = ?$$

$$P(\text{flu} | \text{headache}) = ?$$

$$\begin{aligned} P(\text{headache} \wedge \text{flu}) &= P(\text{headache} | \text{flu}) \times P(\text{flu}) \\ &= 1/2 \times 1/40 = 0.0125 \end{aligned}$$

$$\begin{aligned} P(\text{flu} | \text{headache}) &= P(\text{headache} \wedge \text{flu}) / P(\text{headache}) \\ &= 0.0125 / 0.1 = 0.125 \end{aligned}$$

Bayes Theorem

# Bayes' Rule

$$P(A | B) = \frac{P(B | A) \times P(A)}{P(B)}$$

- Exactly the process we just used
- The most important formula in probabilistic machine learning

(Super Easy) Derivation:

$$P(A \wedge B) = P(A | B) \times P(B)$$

$$P(B \wedge A) = P(B | A) \times P(A)$$

these are the same

Just set equal...

$$P(A | B) \times P(B) = P(B | A) \times P(A)$$

and solve...



**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370-418

# Multi-Value Random Variable

- Suppose  $A$  can take on more than 2 values
- $A$  is a *random variable with arity  $k$*  if it can take on exactly one value out of  $\{v_1, v_2, \dots, v_k\}$
- Thus...

$$P(A = v_i \wedge A = v_j) = 0 \quad \text{if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

$$1 = \sum_{i=1}^k P(A = v_i)$$

A: Month of the Year

EXAMPLE

$$P(A = Jan) = \frac{31}{365} \quad P(A = Feb) = \frac{28}{365}$$

# Marginalization

- We can also show that:

$$P(B) = P(B \wedge [A = v_1 \vee A = v_2 \vee \dots \vee A = v_k])$$

$$P(B) = \sum_{i=1}^k P(B \wedge A = v_i) = \sum_{i=1}^k P(B | A = v_i) P(A = v_i)$$

- This is called **marginalization** over  $A$

**EXAMPLE**     A: Month of the Year; B: Tomorrow is sunny

$$P(\text{Sunny}) = \sum_{i=1}^{12} P(\text{Sunny} | A = \text{Month } i) P(A = \text{Month } i)$$

# Linear algebra review

# Resources

- Zico Kolter, [Linear algebra review](#)
- Sam Roweis's [linear algebra review](#)
- Books:
  - O. Bretscher, Linear Algebra with Applications

# Vectors and matrices

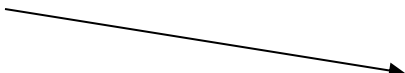
- **Vector** in  $\mathbb{R}^n$  is an ordered set of  $n$  real numbers.

- e.g.  $v = (1, 6, 3, 4)$  is in  $\mathbb{R}^4$

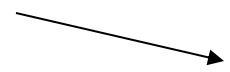
- A column vector:


$$\begin{pmatrix} 1 \\ 6 \\ 3 \\ 4 \end{pmatrix}$$

- A row vector:


$$(1 \ 6 \ 3 \ 4)$$

- $m$ -by- $n$  **matrix** is an object in  $\mathbb{R}^{m \times n}$  with  $m$  rows and  $n$  columns, each entry filled with a (typically) real number:


$$\begin{pmatrix} 1 & 2 & 8 \\ 4 & 78 & 6 \\ 9 & 3 & 2 \end{pmatrix}$$



# Vector operations

- Addition component by component

$$[a_1, a_2, \dots, a_n] + [b_1, b_2, \dots, b_n] = [a_1 + b_1, \dots, a_n + b_n]$$

$$[1, -2, 5] + [0, 3, 7] =$$

- Subtraction is also done component by component

$$[a_1, a_2, \dots, a_n] - [b_1, b_2, \dots, b_n] = [a_1 - b_1, \dots, a_n - b_n]$$

– Can add and subtract row or column vectors of same dimension

- Dot product

– Only works for row and column vector of same size

$$[a_1, a_2, \dots, a_n] \cdot \begin{bmatrix} b_1 \\ \dots \\ b_n \end{bmatrix} = [a_1 b_1, \dots, a_n b_n]$$

$$[1, -2, 5] \cdot \begin{bmatrix} 0 \\ 3 \\ 7 \end{bmatrix} =$$

# Matrix Operations

# Matrix multiplication

We will use upper case letters for matrices. The elements are referred by  $A_{i,j}$ .

- **Matrix product:**

$$A \in \mathbb{R}^{m \times n} \quad B \in \mathbb{R}^{n \times p}$$

$$C = AB \in \mathbb{R}^{m \times p}$$

$$C_{ij} = \sum_{k=1}^n A_{ik} B_{kj}$$

e.g.

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

# Matrix transpose

**Transpose:** You can think of it as

– “flipping” the rows and columns

OR

– “reflecting” vector/matrix on line

e.g.  $\begin{pmatrix} a \\ b \end{pmatrix}^T = (a \quad b)$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

- $(A^T)^T = A$

- $(AB)^T = B^T A^T$

- $(A + B)^T = A^T + B^T$

$A$  is a **symmetric matrix** if  $A = A^T$

# Rank of a Matrix

- $\text{rank}(A)$  (the rank of a  $m$ -by- $n$  matrix  $A$ ) is
  - The maximal number of linearly independent columns
  - The maximal number of linearly independent rows

- If  $A$  is  $n$  by  $m$ , then
  - $\text{rank}(A) \leq \min(m, n)$

- Examples  $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$   $\begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$   $\begin{pmatrix} 2 & 1 & 3 \\ 0 & 5 & 2 \end{pmatrix}$

# Inverse of a matrix

- Inverse of a square matrix  $A$ , denoted by  $A^{-1}$  is the *unique* matrix s.t.
  - $AA^{-1} = A^{-1}A = I$  (identity matrix)
- Inverse of a square matrix exists only if the matrix is **full rank**
- If  $A^{-1}$  and  $B^{-1}$  exist, then
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^T)^{-1} = (A^{-1})^T$

# Diagonal matrices

# System of linear equations

$$\begin{array}{rclcl} 4x_1 & - & 5x_2 & = & -13 \\ -2x_1 & + & 3x_2 & = & 9. \end{array}$$

Matrix formulation

$$Ax = b$$

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}.$$

If  $A$  has an inverse, solution is  $x = A^{-1}b$